Speech Communication 101 (2018) 57–69

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Speech enhancement in spectral envelop and details subspaces

Pengfei Sun[a], Ali Mahdi[a], Jianhong Xu[b], Jun Qin[*,a]

[a] Department of Electrical & Computer Engineering, Southern Illinois University, Carbondale, IL 62901, USA
[b] Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA

ARTICLE INFO

Keywords:
Speech enhancement
Spectral frequency modulation
Low-rank and sparse decomposition
Speech intelligibility

ABSTRACT

Supervised speech enhancement techniques have been proved to improve speech intelligibility. However, one major challenge of supervised approaches involves the overlapped spectral bases between speech and noise components in spectral dictionary space. In this study, we address this challenge through a combination strategy of spectral modulation decoupling and low-rank and sparsity oriented decomposition. Specifically, supervised low-rank and sparse decompositions with energy thresholding are developed in the spectral envelop subspace, In the spectral details subspace, an unsupervised robust principal component analysis is utilized to extract the fine structure. The validation results show that, compared with five speech enhancement algorithms, including MMSE-SPP, NMF-RPCA, RPCA, LARC and BNMF, the proposed algorithms achieves satisfactory performance on improving both perceptual quality and speech intelligibility.

1. Introduction

Speech enhancement is an important topic in speech processing and front end of speech recognition (Sun and Qin, 2017). In general, the goal of speech enhancement is to improve both perceived quality and speech intelligibility by reducing residual noise while minimizing speech signal distortion. Speech with good quality is more comfortable for audiences, while higher speech intelligibility is measured by lower word error rates in speech recognition scenarios. Compared with speech quality, speech intelligibility enhancement is more challenge due to requirement of no distorting of the underlying target speech signal.

Speech intelligibility is mainly affected by vocal tracts, which are generally translated as the spectral envelop (Schroeder and Atal, 1985). The fact that Mel-frequency cepstrum coefficients (MFCC) derived from spectral envelop demonstrates efficacy in automatic speech recognition (ASR) algorithms (Hinton et al., 2012), and further reflects the importance of spectral envelop in terms of speech intelligibility. Generally, the human speech X is a convolution acoustic procedure in the time domain, corresponding to the fact that vocal excitation (harmonics) is modulated by vocal tract (formants) in the frequency domain:

$$X = X_e \circ X_d \tag{1}$$

where the envelop X_e modulates the fine-structure X_d, and $\circ$ is the Hadamard product. Similarly, we can decompose the noisy speech Y as $Y_e \circ Y_d$, where Y_e is the envelop matrix and Y_d is the details matrix. This concept of decoupling frequency modulation has been utilized in previous studies (Ozerov et al., 2012). Simsekli et al. proposed a dynamical source-filter system to incorporate both filter (i.e., envelop) and excitation (i.e., fine-structure) dictionaries into the proposed statistical model (Simsekli et al., 2014). This algorithm implemented maximum a-posteriori estimation by jointly updating the speech components in two subspaces. Decoupling frequency modulation intrinsically reflects speech producing process, and recovering speech components separately in envelop and details subspaces may reduce the overlap ratio between speech and noise.

Low-rank and sparse decomposition (LSD) demonstrates good performance for speech denoising in several works (Duan et al., 2012; Sun and Qin, 2016). This category of methods is proved to accommodate nonstationary background noise because the activation of noise components can be temporally variable. In addition, supervised LSD inherits the merits of dictionary based non-negative matrix factorization (NMF) technique (Mohammadiha et al., 2013) and sparse coding approach (Sigg et al., 2012), and at the meantime LSD related techniques can avoid the determination of the rank of noise in advance. However, a common issue for LSD based speech enhancement approach is that speech bases may highly overlap with the convex hull of noise bases in the spectrum domain. Another weakness of LSD method comes from the fact that speech and noise often demonstrate both low-rank and sparse properties.

In this study, we propose a modulation decoupling based algorithm combining dictionary and non-dictionary LSDs. The proposed method inherits the merits of both frequency decoupling and LSD. In the envelop subspace Y_e, driven by the motivation of improving intelligibility and consideration of sparsity-to-low-rank ratio (SLR), two specifically

* Corresponding author.
E-mail address: jqin@siu.edu (J. Qin).

https://doi.org/10.1016/j.specom.2018.05.006
Received 22 November 2017; Received in revised form 17 May 2018; Accepted 25 May 2018
Available online 07 June 2018
0167-6393/

designed algorithms, referring as two-layer LSD (TLSD) and single-layer LSD (SLSD), are comparatively introduced to implement the speech extraction. An offline trained speech envelop dictionary is utilized in both TLSD and SLSD algorithms. In the detail subspace Y_d, a general unsupervised LSD is used to obtain speech components X_d. The spectrogram of estimated speech can be obtained as the element-wise product of the two extracted submatrices.

1.1. Related work

In this study, our proposed speech enhancement algorithm can be categorized as modulation decoupling based LSD. By exploiting intrinsic decomposition through convex optimization, low-rank and sparsity analysis overcomes the high sensitivity of the conventional principal component analysis (PCA) when subjecting to large corruptions. Modulation based source separation technologies are developed based on the knowledge that the spectrogram of speech can be described by a time-varying weighted sum of component modulations as well as spectral frequency modulation (Elliott and Theunissen, 2009).

1.1.1. Low-rank and sparse decomposition

The idea of applying LSD to separate speech from background noise is derived from the intrinsic data structure of noisy speech spectrogram (Sun and Qin, 2016), in which background noise usually demonstrates low spectral diversity whereas speeches are more instantaneous and changeable. Specific constraints (e.g., masking threshold, noise rank, and block-wise restrictions) (Yang, 2012; Sun et al., 2014) are incorporated to optimize the decomposition. LSD has been implemented in the wavelet packet transform domain (Röbel et al., 2007; Bouzid et al., 2016), in which the speech components are concentrated to exhibit more sparsity.

In many relevant cases, using a single spectral model to describe the speech signal is insufficient. Because with long-term repeated structure, speech can also demonstrate low-rank characteristic as well as sparsity. The coexistence of low-rank and sparse properties in speech requires a more comprehensive constraint to reflect its spectral structure. Chen and Ellis (2013) utilized a modified robust PCA (RPCA) optimization function, in which offline trained speech spectral dictionary is employed and outlying entries are subjected to minimal energy restriction. Duan et al. introduced an online learned dictionary to implement non-negative spectrogram decomposition (Duan et al., 2012). Yang proposed an LSD strategy via combining dictionaries with respect to speech and noise (Yang, 2013).

1.1.2. Modulation based source separation

In speech spectrogram, time and frequency modulations are intuitively represented as correlations among neighboring spectral magnitudes. These correlations have been frequently employed as a prior knowledge to improve either the noise power estimation (Gerkmann and Hendriks, 2012) or speech magnitude estimation (Cohen, 2003). Typically, by incorporating 1D smooth coefficients (Martin, 2001) or 2D average window (Gerkmann et al., 2008) imposed on the spectrogram, significant improvements on speech quality can be achieved by taking correlations into account.

Instead of locally introducing correlations derived from the modulations in spectrogram, a straightforward method is to decouple modulation. By utilizing pseudo frequencies, Deng et al. conducted a conditional minimum mean square error (MMSE) estimation in the cepstrum domain, and the result showed that it was a noise-robust feature selection approach (Deng et al., 2004). Clark and Atlas proposed a generalized coherent modulation filter to recover arbitrarily chosen envelope (Clark and Atlas, 2009). Their work (Clark et al., 2011) further indicates that demodulation based low frequency envelop structure analysis can potentially promote the speech recognition performances. Veisi and Sameti introduced hidden markov models into the mel-frequency domain (Veisi and Sameti, 2013), and results indicated a significant improvement on noise cancellation. Different from frequency modulation algorithms, Paliwal et al. proposed a frame-wise transformation along the time axis, and clean speech is obtained based on conventional speech estimator (Paliwal et al., 2012). Factoring modulation strategy has been widely used to model speech signals and eliminate the noise in corresponding subspaces (Kollmeier and Koch, 1994; Durrieu et al., 2010; Ozerov et al., 2012).

1.2. Method overview

To obtain the proposed two modulation subspaces, a cepstrum based modulation inverse (CMI) transform is applied. It first obtains cepstrogram by applying element-wise logarithm and discrete Fourier transform (DFT), and then window functions are used to separate the envelop and details subspaces in the cepstrum domain. Finally, inverse Fourier transform is implemented to obtain two modulation subspaces (Simsekli et al., 2014).

In each subspace, LSD is implemented to extract the speech components. Considering that the spectral envelop subspace has a slowly varying property, noise components in this subspace share more spectral bases with speech components than those in the spectral details subspace. Therefore, in the spectral envelop subspace, supervised LSD can be applied, in which two different decomposition strategies adapting to different types of noises are proposed. In the spectral details subspace Y_d, the speech components show highly regular structure (i.e., fine structure), and noise is supposed to be low-rank. A typical unsupervised RPCA method can be used to effectively extract the speech spectral details. Specifically, for unvoiced sounds, the supervised LSD in the envelop subspace determines the magnitude of the speech components. Therefore, even fricatives in the details subspace show similar structures as noises, the accurate recovery of spectral envelop can greatly reduce impact of noise from the details subspace. Moreover, the harmonics components in speech can be concentrated, which leads to better separation results comparing with the general spectrogram decomposition. The implementation procedure is shown in Fig. 1.

1.3. Contribution of our work

By decoupling the spectral envelop and details subspaces, LSD is implemented in both subspaces. The contributions of this study can be summarized as follows:

- We propose a spectral frequency modulation based speech

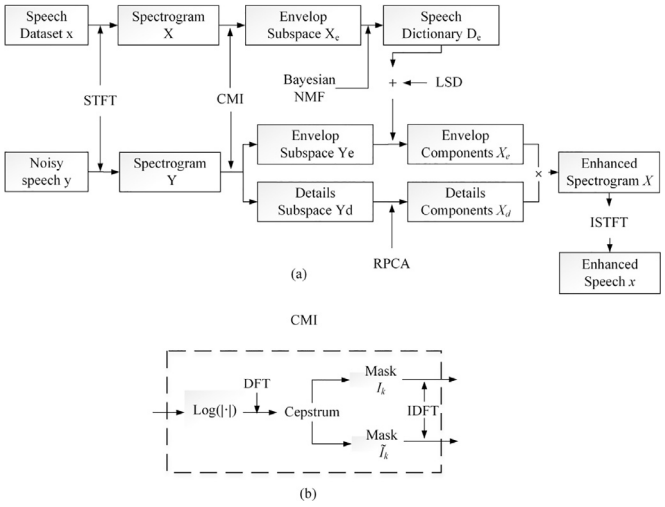

Fig. 1. The schematic diagram of (a) the proposed low-rank and sparse decomposition in the modulation based subspaces and (b) details of the proposed CMI procedure.